DATA LAKES-A NEW DATA REPOSITORY FOR BIG DATA ANALYTICS WORKLOADS

Article *in* International Journal of Advanced Computer Research · October 2016

CITATIONS
2

READS
632

2 authors, including:

Divya Meena Sundaram
VIT University
27 PUBLICATIONS **30** CITATIONS

SEE PROFILE

ISSN No. 0976-5697

Volume 7, No. 5, September-October 2016

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

DATA LAKES – A NEW DATA REPOSITORY FOR BIG DATA ANALYTICS WORKLOADS

Ms. S. Divya Meena
Assistant Professor,
Department of Computer Science,
Jansons Institute of Technology, Coimbatore, India

Ms. S. Vidhya Meena
BE-CSE,
Kingston Engineering College,
Vellore 632059, India

Abstract Today, 67% of all storage capability support unstructured information. By 2017, this can increase to 80%. Firm's unit understand that to resolve these growth challenges an approach is needed to manage and maintain the data. An info lake or knowledge hub is also a scalable infrastructure that is economically partaking and designed for flexibility [1]. It provides developers one place to store all of their structured and semi-structured information in its native format whereas not having to stress regarding storage and capability limitations on individual files. In 2015, information lakes will evolve as organizations move from batch to data processing and integrate file-based Hadoop and information engines into their large-scale processing platforms. The big trend in 2015 goes to be around the continuous access and method of events and information in real time to appreciate constant awareness and take immediate action.

***Keywords*:** Information Lake, information Warehouse, huge information Analytics, Hadoop.

I. INTRODUCTION

Wherever the data lake style is live, and as that information lake becomes wider and deeper; the addition of an information repository will augment a bunch of capricious knowledge objects in a very method that is more intelligible at people-scale. Though an info lake is not gift, it specialize in the creation of a comprehensive knowledge repository that still has edges to the enterprise in terms of enlarged standardization, information integrity, modification management, and owner answerability [2]. Knowledge has the power to make all information, yet offer, out there and perceivable to any or all or any information shoppers, even so of the role, even so of the scale of the data lake.

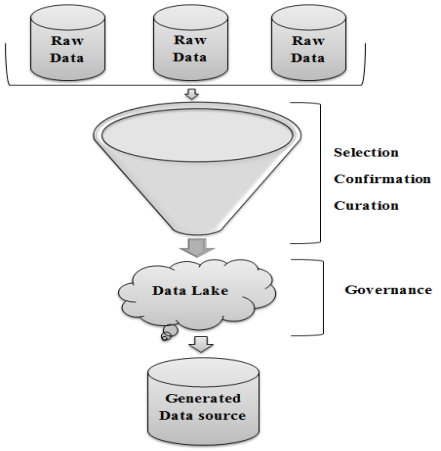
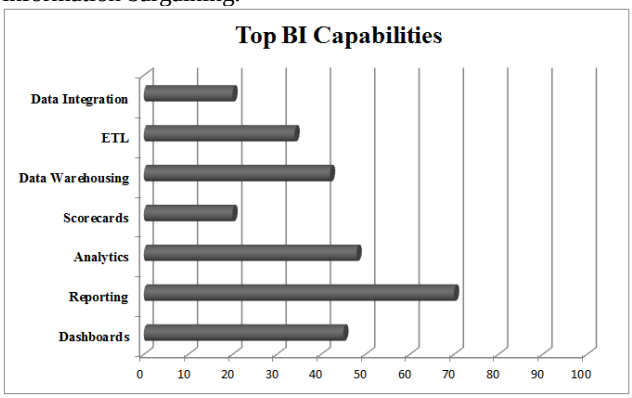


Fig 1: Knowledge bargaining method

At the tip of the day, the data shopper does not care where the data is housed. Their real issues centre on the provision of the data they need, and conjointly the fit-for-use quality of that information [3]. Over time, information has less worth and extra risk associated with it. It does not be to still fill the lake whereas not some attempt to drain off the data that has become noisier. during this paper, we tend to gift variety of challenges inherent in making, filling, maintaining, and governing a curated knowledge lake, a group of processes that conjointly outline the actions of information bargaining.



The higher than diagram will be in brief mentioned as knowledge lake method involving the subsequent steps [4];

1) ***Load or archive batch data:*** Knowledge is loaded from completely different sources like transactions, social media, web logs, machine devices et al.
2) ***Replicate modified knowledge and schemas:*** Some documents tend to follow schemas that aren't meant for Knowledge Lake. Such documents square measure replicated and also the schemas square measure modified,
3) ***Stream period of time knowledge:*** These data aren't restricted to social media, web logs, sensors, machine devices, etc.
4) ***Mask sensitive data:***Naturally, the sensitive knowledge square measure protected by means that of various masking techniques.
5) ***Access client golden record:*** Golden record indicates the info from individuals; they're named therefore to indicate the worth of the info the client holds.

6) ***Refine and reverend knowledge:*** Curation involves many steps needed to keep up and utilize digital data throughout its life-cycle for current and future interested users.
7) ***Move results to enterprise knowledge warehouse:*** The raw supply is refined to a lot of helpful knowledge and also the knowledge is then enraptured to data warehouse for storage.
8) ***Explore and validate knowledge:*** The generated knowledge is valid by analytics method; discovering any potential issues within the data and news the result. The result's then ensured if legal conformity is satisfactory.
9) ***Govern and enrich with metadata:*** Information can enrich the generated knowledge by adding additional essential info concerning the info hold on in warehouse.
10) ***Correlate period of time events:*** The period of time events square measure related to with historical patterns and trends.

II. WHAT IS KNOWLEDGE LAKE AND WHY IS IT WIDELY HELD

The idea of a knowledge lake is rising as well-liked; thanks to the organization and building of consequent generation of systems to master new massive data challenges [5]. It has become fashionable as a result of it providing cheap and technologically possible massive knowledge challenges. Organizations square measure discovering the knowledge lake as an evolution from their existing data design.

Doles of Knowledge Lake:

1) A very scalable, economical infrastructure that lowers costs and easily keeps pace with growing information storage requirements
2) Powerful, easy-to-use analytic tools unlock the business worth of informationthat lives in data
3) Enterprise class information protection to maximise accessibility and security selections to meet business requirements
4) Customers get the foremost enterprise-grade and simple to use Hadoop huge information package, all with 24/7 world support and services [6].

Benefits of Knowledge Lake in Enterprise:

1) ***Economical Storage:*** Eliminates storage silos, simplifies management, and improves utilization
2) ***Large Scalability:*** Built from scale-out architectures that unit massively scalable and simple to manage
3) ***Enlarged Operational Flexibility:*** Multi-protocol and next-generation access capabilities support ancient and rising applications
4) ***Enterprise Attributes:*** Protects information with economical and resilient backup, disaster recovery and security selections
5) ***In-Place huge information Analytics:*** Leverages shared storage and support for protocols like HDFS to deliver cost-efficient, in-place analytics with faster time to results.

III. DATA LAKE VS KNOWLEDGE WAREHOUSE

The distinction between a knowledge lake and a knowledge warehouse is that in a knowledge warehouse, the data is pre-categorized at the aim of entry, which could dictate but it's getting to be analysed. The matter is that, at intervals the globe of big information, we've an inclination to don't extraordinarily perceive what worth the data has when; it's at the beginning accepted from the array of sources [7]. We would perceive some queries we would wish to answer, but to not the extent that it's smart to shut off the flexibleness to answer queries that happen later. Therefore, storing information in some optimal sort for later analysis doesn't produce any sense. Although Information Lake is every technically and financially come-at-able, operational implementation of big, merely accessible repository supported budget storage, still it might raise further queries than it resolves. Some of the properties of knowledge warehouse are thatit represents associate abstracted image of the business organized by subject, it is extraordinarily reworked and structured and information is not loaded to the data warehouse until the use for it has been printed. Some of the properties of Information Lake are thatall knowledge is loaded from offer systems [8]. No information is turned away, information is hold on at the leaf level in associate untransformed or nearly untransformed state and information is reworked and schema is applied to satisfy the necessities of study [9].

FACET	DATA WAREHOUSE	DATA LAKE
Workload	Batch processing. Enhances query performance. Supports interactive users	Batch processing of data at scale. Currently improving its capabilities to support more interactive users
Schema	Schema on write	Schema on read
Scale	Can scale to large data volumes at moderate cost	Can scale to extreme data volumes at low cost
Access methods	Data accessed through standard SQL and standardized BI tools	Data accessed through programs of SQL-like systems, and other methods
Complexity	Complex join	Complex processing
Data	Cleansed	Raw
Cost/ Efficiency	Efficient use of CPU/IO	Low cost of storage and processing

IV. DATA LAKE MISCONCEPTION

The growing promotion encompassing information lakes is inflicting substantial confusion at regular intervals in the data management house. Several vendors unit promoting information lakes as a significant part try to maximise huge information opportunities [10]. In broad terms, information lakes unit marketed as enterprise-wide information management platforms for analysing disparate sources of

knowledge in its native format. The idea is simple: instead of golf stroke information throughout a purpose-built information store, you progress it into an info lake in its original format. This eliminates the direct costs of knowledge uptake, like transformation. Once information is placed into the lake, it's out there for analysis by everyone at regular intervals in the organization [11]. The need for enlarged lightness and accessibility for information analysis is the first driver for information lakes. Even so, whereas it's positively true that Information Lake can ask price to varied elements of the organization; the proposition of enterprise-wide information management has but to be completed.

Data lakes carry substantial risks. The foremost one is that the inabilities to ascertain information quality or the lineage of findings by different analysts or users that have found worth, previously, in exploitation a similar information at intervals [12]. Another risk is security and access management. Information is commonly placed into the data lake with no oversight of the contents. Finally, performance aspects should not be unnoticed. Tools and information interfaces just cannot perform at a similar level against a general-purpose store as they are going to be against optimized and purpose-built infrastructure [13].

V. CONCLUSION

A well-constructed knowledge repository will modify the enterprise to leap-frog over the data lake and empower the delivery of knowledge as a Service, Analytics as a Service, advanced analytics, Self-service information provisioning and information Science sandbox provisioning [14]. A maturing knowledge program is in addition crucial to resolving the issues around governance, security, and conjointly the management of unstructured and third party information. Knowledge transparency is basic to democratizing and extracting worth from huge information [15].By following the knowledge information, we've shown that the emergence of the info lake comes from the necessity to manage and exploit new sorts of data. In essence, the form of your knowledge lake is decided by what you wish to try to [16]. The proper knowledge lake will solely be created through experimentation. Together, the knowledgelake and also the enterprise data warehouse give a natural process of capabilities that delivers fast returns. Permitting individuals to try a lot of with knowledge gives quicker and driving business results. That’s the last word of payback from finance in every knowledge lake to enrich your enterprise knowledge warehouse.

VI. REFERENCES

[1] Henschen, D. 2011. “Why All the Hadoop?” Information Week, November 14, pp. 19-26.

[2] Lusch, R. F., Liu, Y., and Chen, Y. 2010. “The Phase Transition of Markets and Organizations: The New Intelligence and Entrepreneurial Frontier,” IEEE Intelligent Systems (25:1), pp. 71-75.

[3] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. 2011. “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” McKinsey Global Institute.

[4] Stonebraker, Michael, Daniel Bruckner, Ihab F. Ilyas, George Beskales, Mitch Cherniack, Stanley B. Zdonik, Alexander Pagan, and Shan Xu. "Data Curation at Scale: The Data Tamer System." In CIDR. 2013.

[5] Vassiliadis, Panos. "A survey of Extract–transform–Load technology."International Journal of Data Warehousing and Mining (IJDWM) 5, no. 3 (2009): 1-27.

[6] Haas, L., Cefkin, M., Kieliszewski, C., Plouffe, W., Roth, Mary., The IBM Research Accelerated Discovery Lab. 2014 SIGMOD.

[7] Haas, Laura M., Mauricio A. Hernández, Howard Ho, Lucian Popa, and Mary Roth. "Clio grows up: from research prototype to industrial tool." In Proceedings of the 2005 ACM SIGMOD international conference on Management of data, pp. 805-810. ACM, 2005.

[8] Halbert, Martin. "Prospects for research data management." Research Data Management (2013). 1: http://www.clir.org/pubs/reports/pub160/pub160.pdf

[9] Halevy, Alon, Anand Rajaraman, and Joann Ordille. "Data integration: the teenage years." In Proceedings of the 32nd international conference on Very large data bases, pp. 9-16. VLDB Endowment, 2006.

[10] Hassanzadeh, O., et. al. “Helix: Online Enterprise Data Analytics”. Proceedings of the 20th international conference companion on the World wide.

[11] Adomavicius, G., and Tuzhilin, A. 2005. “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,” IEEE Transactions onKnowledge and Data Engineering (17:6), pp. 734-749.

[12] Chiang, R. H. L., Goes, P., and Stohr, E. A. 2012. “Business Intelligence and Analytics Education and Program Development: A Unique Opportunity for the Information Systems Discipline,” ACM Transactions on Management Information Systems (3:3), forthcoming.

[13] Chen, H., Brandt, L., Gregg, V., Traunmuller, R., McIntosh, A., Dawes, S., Hovy, E., and Larson, C. A. (eds.). 2007. DigitalGovernment: E-Government Research, Case Studies, and Implementation, New York: Springer.

[14] Angevaare, Inge. 2009. Taking Care of Digital Collections and Data: 'Curation' and Organisational Choices for Research Libraries. LIBER Quarterly: The Journal of European Research Libraries19, no. 1 (2009): 1-12.

[15] Chessell, M., Scheepers, F., Nguyen, N., van Kessel, R., and van der Starre, R. 2014. Governing and Managing Big Data for Analytics and Decision Makers. IBM Redguides for Business Leaders.

[16] Halevy, Alon, Anand Rajaraman, and Joann Ordille. "Data integration: the teenage years." In Proceedings of the 32nd international conference on Very large data bases, pp. 9-16. VLDB Endowment, 2006.